University of Waterloo
ECE 657A: Data and Knowledge Modeling and Analysis
Winter 2016
Assignment 2: Classification and Clustering
Due: Friday March 11, 2016 (before midnight)

Assignment Type*:* group, the max team members is 3.
Hand in**:** one report (PDF) per group, via the LEARN dropbox. Also submit the code / scripts needed to reproduce your work.
Objective**:** To gain experience on the use of classification and clustering methods

Data sets (available on the UW 'LEARN' system)

Dataset D (DataD.mat)	This data is the splice junctions on DNA sequences. The given dataset includes 2200 samples with 57 features, in the matrix 'fea'. It is a binary class problem. The class labels are either +1 or -1, given in the vector 'gnd'. Parameter selection and classification tasks are conducted on this dataset.
Dataset F (DataF.mat)	This is a handwritten collection including digits 0 to 9. The given dataset includes 6200 samples with 256 features, given in the matrix 'fea'. This dataset is used in clustering tasks. The sample class labels ('gnd') are used for the purpose of performance evaluation.

I. Parameter Selection and Classification (for dataset D)

Classify dataset D using five classifiers: k-NN, Support Vector Machine (with RBF kernel), Naïve Bayes Classifier, Decision Trees and Neural Networks. The objective is to experiment with parameter selection in training classifiers and to compare the performance of these well- known classification methods.

1) Preprocess the given data using the Z-score normalization, and split the data into two halves, the first half being the training set and the second half being the test set. (Normally you would do this randomly, but for this assignment a deterministic split will make the rest of the answers easier to grade).
2) For k-NN you need to evaluate the best value **k** to use. Using 5-fold cross validation (the `crossvalind` function can help) on the training set evaluate k-NN on the values k=[1, 3, 5, 7, …, 31]. Plot a figure that shows the relationship between the accuracy and the parameter **k**. Report the best **k** in terms of classification accuracy.

3) For the RBF kernel SVM, there are two parameters to be decided: the soft margin penalty term "**c**" and the kernel width parameter "**gamma**". Again use 5-fold cross validation on the training set to select the parameter "**c**" from the set [0.1, 0.5, 1, 2, 5, 10, 20, 50] and select the parameter "**gamma**" from the set [0.01, 0.05, 0.1, 0.5, 1, 2, 5, 10]. Report the best parameters in terms of classification accuracy including plotting the ROC curves.

4) Using the chosen parameters from the above parameter selection process for k-NN and SVM, and the default setups for Naïve Bayes classifier, Decision Tree and Neural Network, classify the test set. Repeat each classification method 20 times by varying the split of training-test set (now select a random half of the data for training and the other half for test). Report the average and standard deviation of classification performance on the test set regarding: accuracy, precision, recall, and F- Measure. Also report the training time and classification time of the four methods.

5) Comment on the obtained results, what are the benefits and weaknesses of each method on this dataset. How could this analysis help to make the choice of the right method to use for a dataset of this type in the future?

II. Clustering Analysis (for dataset F)

The data has already been normalized into the range of [-1, 1], the sample labels are used for the purpose of performance evaluation.

Apply PCA to reduce the dimension to be 4, and then conduct the following clustering analysis.

1) Perform hierarchical clustering using agglomerative algorithms:
   a) Stop when the number of clusters is 10 (the same as the number of given classes). Compare the linkage methods of "single", "complete", and "ward (minimum variance algorithm)". Evaluate the clustering results in terms of Separation-Index, Rand-Index, and F-measure. Compare the three linkage types and comment.
   b) Fix the linkage type as "ward", study the number of clusters from 2 to 15, increment by 1, what is the optimal number of clusters suggested by Separation- Index in this case?
2) Cluster the data using k-means algorithm.
   a) Run the algorithm for the number of clusters k from 2 to 15, increment by 1. Evaluate the clustering results in terms of Separation-Index, Rand-Index, and F-measure.
   b) Plot these evaluation measures with respect to the number of clusters. What is the optimal number of clusters suggested by these indexes?
3) Cluster the data using fuzzy c-means algorithm, with number of clusters as 10, and the fuzzy parameter (the exponent for partition matrix) m = 2.
   a) Plot the average cluster membership values for the samples of the digit '1' and '3'; explore their overlaps with other clusters.
   b) If we produce hard clustering by making the max membership value of a sample to be one and the rest of its membership values zero, evaluate the clustering results. Comment on the obtained results.

Note:

1. For classification methods of k-NN, Naive Bayes, SVM, and Decision tree, Matlab has implemented functions as: `knnclassify`, `NaiveBayes.fit/predict`, `svmtrain /svmclassify`, and `classregtree /eval`.

2. There is a well-known SVM library named libSVM, which is implemented in C++ and includes a matlab interface (`http://www.csie.ntu.edu.tw/~cjlin/libsvm/`)

3. For the clustering methods of k-means, fuzzy c-means, and hierarchical clustering, Matlab has implementation as `kmeans`, `fcm`, and `linkage/clusterdata`.

4. For Neural Networks you should be able to use the Matlab `patternnet` function. See (`http://www.mathworks.com/help/nnet/gs/classify-patterns-with-a-neural-network.html`) for a tutorial on setting up neural networks visually and generating the initial code.

5. Late submissions (up to 3 days) are accepted with penalty of 10% per day.